Product and Customer Segment Analysis

Phuong Tran

In this document, transactiondata and purchasedata were processed and explored:

1. Examined transactiondata:

- Checked for outliers and missing values
- Added extra features including PACK_SIZE and BRAND_NAME

2. Examined purchasedata:

- Checked for nulls
- Checked for distribution of customers based on LIFESTAGE and PREMIUM_GROUPS

3. Merged transactiondata and purchasedata for analysis:

- Explored which customer segments drove total sales, product quantity and product price
- Performed t-test to confirm the significance of difference
- Explored which Brands were preferred by each Customer Segment, visualized with mosaic plot and significance tested with Pearson Chi-square test
- Explored which PACK_SIZEs were preferred by each Customer Segment, visualized with mosaic plot and significance tested with Pearson Chi-square test

```r
# Load packages
library(data.table)
library(ggplot2)
library(readxl)
library(readr)
library(dplyr)
library(tidyr)
library(arules)
library(methods)
library(ggmosaic)
```

```r
# Import data
transactiondata <- read_excel("QVI_transaction_data.xlsx")
purchasebehaviour <- read_csv("QVI_purchase_behaviour.csv")
```

```r
# Examine transaction data
head(transactiondata)
```

```
## # A tibble: 6 x 8
##    DATE STORE_NBR LYLTY_CARD_NBR TXN_ID PROD_NBR PROD_NAME    PROD_QTY TOT_SALES
##   <dbl>     <dbl>          <dbl>  <dbl>    <dbl> <chr>           <dbl>     <dbl>
## 1 43390         1           1000      1        5 Natural Chi~        2       6
## 2 43599         1           1307    348       66 CCs Nacho C~        3       6.3
## 3 43605         1           1343    383       61 Smiths Crin~        2       2.9
## 4 43329         2           2373    974       69 Smiths Chip~        5      15
## 5 43330         2           2426   1038      108 Kettle Tort~        3      13.8
## 6 43604         4           4074   2982       57 Old El Paso~        1       5.1
```

```
# Convert DATE column to date format
transactiondata$DATE<-as.Date(transactiondata$DATE,origin = "1899-12-30")
head(transactiondata)
```

```
## # A tibble: 6 x 8
##   DATE       STORE_NBR LYLTY_CARD_NBR TXN_ID PROD_NBR PROD_NAME PROD_QTY
##   <date>         <dbl>          <dbl>  <dbl>    <dbl> <chr>        <dbl>
## 1 2018-10-17         1           1000      1        5 Natural ~        2
## 2 2019-05-14         1           1307    348       66 CCs Nach~        3
## 3 2019-05-20         1           1343    383       61 Smiths C~        2
## 4 2018-08-17         2           2373    974       69 Smiths C~        5
## 5 2018-08-18         2           2426   1038      108 Kettle T~        3
## 6 2019-05-19         4           4074   2982       57 Old El P~        1
## # ... with 1 more variable: TOT_SALES <dbl>
```

```
#Summary of PROD_NAME
table(transactiondata$PROD_NAME,useNA = "ifany") # no NA values
```

```
## 
##                         Burger Rings 220g 
##                                      1564 
##                  CCs Nacho Cheese    175g 
##                                      1498 
##                         CCs Original 175g 
##                                      1514 
##                  CCs Tasty Cheese    175g 
##                                      1539 
##            Cheetos Chs & Bacon Balls 190g 
##                                      1479 
##                        Cheetos Puffs 165g 
##                                      1448 
##                      Cheezels Cheese 330g 
##                                      3149 
##                  Cheezels Cheese Box 125g 
##                                      1454 
##            Cobs Popd Sea Salt  Chips 110g 
##                                      3265 
##    Cobs Popd Sour Crm  &Chives Chips 110g 
##                                      3159 
## Cobs Popd Swt/Chlli &Sr/Cream Chips 110g 
##                                      3269 
##          Dorito Corn Chp     Supreme 380g 
##                                      3185 
##         Doritos Cheese      Supreme 330g
##                                      3052
##  Doritos Corn Chip Mexican Jalapeno 150g
##                                      3204
##  Doritos Corn Chip Southern Chicken 150g
##                                      3172
##  Doritos Corn Chips  Cheese Supreme 170g
##                                      3217
##    Doritos Corn Chips  Nacho Cheese 170g
##                                      3160
##        Doritos Corn Chips  Original 170g
##                                      3121
##                 Doritos Mexicana    170g
##                                      3115
##          Doritos Salsa       Medium 300g
##                                      1449
##                 Doritos Salsa Mild  300g
##                                      1472
##           French Fries Potato Chips 175g
##                                      1418
##    Grain Waves         Sweet Chilli 210g
##                                      3167
##    Grain Waves Sour    Cream&Chives 210G
##                                      3105
##    GrnWves Plus Btroot & Chilli Jam 180g
##                                      1468
##  Infuzions BBQ Rib   Prawn Crackers 110g
##                                      3174
##  Infuzions Mango     Chutny Papadums 70g
##                                      1507
## Infuzions SourCream&Herbs Veg Strws 110g
##                                      3134
## Infuzions Thai SweetChili PotatoMix 110g
##                                      3242
##   Infzns Crn Crnchers Tangy Gcamole 110g
##                                      3144
##             Kettle 135g Swt Pot Sea Salt
##                                      3257
##                       Kettle Chilli 175g
##                                      3038
##         Kettle Honey Soy    Chicken 175g
##                                      3148
##   Kettle Mozzarella   Basil & Pesto 175g
##                                      3304
##                     Kettle Original 175g
##                                      3159
##     Kettle Sea Salt     And Vinegar 175g
##                                      3173
##       Kettle Sensations   BBQ&Maple 150g
##                                      3083
## Kettle Sensations   Camembert & Fig 150g
##                                      3219
##    Kettle Sensations   Siracha Lime 150g
##                                      3127
##  Kettle Sweet Chilli And Sour Cream 175g
##                                     3200
##  Kettle Tortilla ChpsBtroot&Ricotta 150g
##                                     3146
##     Kettle Tortilla ChpsFeta&Garlic 150g
##                                     3138
## Kettle Tortilla ChpsHny&Jlpno Chili 150g
##                                     3296
##   Natural Chip        Compny SeaSalt175g
##                                     1468
##  Natural Chip Co     Tmato Hrb&Spce 175g
##                                     1572
##   Natural ChipCo      Hony Soy Chckn175g
##                                     1460
##   Natural ChipCo Sea  Salt & Vinegr 175g
##                                     1550
##   NCC Sour Cream &    Garden Chives 175g
##                                     1419
## Old El Paso Salsa   Dip Chnky Tom Ht300g
##                                     3125
##  Old El Paso Salsa   Dip Tomato Med 300g
##                                     3114
## Old El Paso Salsa   Dip Tomato Mild 300g
##                                     3085
##                 Pringles Barbeque   134g
##                                     3210
##      Pringles Chicken    Salt Crips 134g
##                                     3104
##         Pringles Mystery    Flavour 134g
##                                     3114
##          Pringles Original   Crisps 134g
##                                     3157
##                 Pringles Slt Vingar 134g
##                                     3095
##           Pringles SourCream  Onion 134g
##                                     3162
##         Pringles Sthrn FriedChicken 134g
##                                     3083
##             Pringles Sweet&Spcy BBQ 134g
##                                     3177
##   Red Rock Deli Chikn&Garlic Aioli 150g
##                                     1434
##  Red Rock Deli Sp    Salt & Truffle 150G
##                                     1498
## Red Rock Deli SR    Salsa & Mzzrlla 150g
##                                     1458
##     Red Rock Deli Thai  Chilli&Lime 150g
##                                     1495
##         RRD Chilli&         Coconut 150g
##                                     1506
##         RRD Honey Soy       Chicken 165g
##                                     1513
##                 RRD Lime & Pepper   165g
##                                     1473
##                    RRD Pc Sea Salt     165g 
##                                        1431 
##                    RRD Salt & Vinegar  165g 
##                                        1474 
##         RRD SR Slow Rst     Pork Belly 150g 
##                                        1526 
##        RRD Steak &         Chimuchurri 150g 
##                                        1455 
##         RRD Sweet Chilli &  Sour Cream 165g 
##                                        1516 
##          Smith Crinkle Cut   Bolognese 150g 
##                                        1451 
##       Smith Crinkle Cut   Mac N Cheese 150g 
##                                        1512 
##       Smiths Chip Thinly  Cut Original 175g 
##                                        1614 
##      Smiths Chip Thinly  CutSalt/Vinegr175g 
##                                        1440 
##      Smiths Chip Thinly  S/Cream&Onion 175g 
##                                        1473 
##           Smiths Crinkle      Original 330g 
##                                        3142 
##    Smiths Crinkle Chips Salt & Vinegar 330g 
##                                        3197 
##     Smiths Crinkle Cut  Chips Barbecue 170g 
##                                        1489 
##      Smiths Crinkle Cut  Chips Chicken 170g 
##                                        1484 
##     Smiths Crinkle Cut  Chips Chs&Onion170g 
##                                        1481 
##     Smiths Crinkle Cut  Chips Original 170g 
##                                        1461 
##    Smiths Crinkle Cut  French OnionDip 150g 
##                                        1438 
##     Smiths Crinkle Cut  Salt & Vinegar 170g 
##                                        1455 
##         Smiths Crinkle Cut  Snag&Sauce 150g 
##                                        1503 
##       Smiths Crinkle Cut  Tomato Salsa 150g 
##                                        1470 
##      Smiths Crnkle Chip  Orgnl Big Bag 380g 
##                                        3233 
##    Smiths Thinly       Swt Chli&S/Cream175G 
##                                        1461 
##      Smiths Thinly Cut   Roast Chicken 175g 
##                                        1519 
##        Snbts Whlgrn Crisps Cheddr&Mstrd 90g 
##                                        1576 
##    Sunbites Whlegrn    Crisps Frch/Onin 90g 
##                                        1432 
##      Thins Chips         Originl saltd 175g 
##                                        1441 
##              Thins Chips Light&  Tangy 175g 
##                                        3188
##         Thins Chips Salt &  Vinegar 175g
##                                     3103
##         Thins Chips Seasonedchicken 175g
##                                     3114
##     Thins Potato Chips  Hot & Spicy 175g
##                                     3229
##          Tostitos Lightly    Salted 175g
##                                     3074
##        Tostitos Smoked     Chipotle 175g
##                                     3145
##            Tostitos Splash Of  Lime 175g
##                                     3252
##                 Twisties Cheese     270g
##                                     3115
##          Twisties Cheese     Burger 250g
##                                     3169
##                     Twisties Chicken270g
##                                     3170
##   Tyrrells Crisps     Ched & Chives 165g
##                                     3268
##  Tyrrells Crisps     Lightly Salted 165g
##                                     3174
##           Woolworths Cheese   Rings 190g
##                                     1516
##           Woolworths Medium   Salsa 300g
##                                     1430
##           Woolworths Mild     Salsa 300g
##                                     1491
##         WW Crinkle Cut      Chicken 175g
##                                     1467
##        WW Crinkle Cut      Original 175g
##                                     1410
##        WW D/Style Chip     Sea Salt 200g
##                                     1469
##           WW Original Corn    Chips 200g
##                                     1495
##           WW Original Stacked Chips 160g
##                                     1487
##   WW Sour Cream &OnionStacked Chips 160g
##                                     1483
##      WW Supreme Cheese   Corn Chips 200g
##                                     1509
```

```r
n_distinct(transactiondata$PROD_NAME) #114 distinct product names
```

```
## [1] 114
```

```r
# Split PROD_NAME entries to words by space and then rename the column to words
productWords<-data.table(unlist(strsplit(unique(transactiondata$PROD_NAME)," ")))
productWords<-setNames(productWords,"words")
```

```r
# Clean productWords from blank rows digits and special characters
productWords<-productWords[!(words=="&"|words=="")][-grep("^[0-9]",words)]
```

```r
# Find common words among PROD_NAME
freq_words<-as.data.frame(table(productWords))

#Top 20 most common words
head(freq_words[order(-freq_words$Freq),],20) # order descending according to Frequency
```

```
##     productWords Freq
## 39         Chips   21
## 151       Smiths   16
## 58       Crinkle   14
## 65           Cut   14
## 92        Kettle   13
## 25        Cheese   12
## 140         Salt   12
## 115     Original   10
## 36          Chip    9
## 71       Doritos    9
## 139        Salsa    9
## 54          Corn    8
## 129     Pringles    8
## 136          RRD    8
## 28       Chicken    7
## 196           WW    7
## 143          Sea    6
## 155         Sour    6
## 32        Chilli    5
## 60        Crisps    5
```

```r
# Remove Salsa entries
transactiondata<-transactiondata[grep("SALSA",transactiondata$PROD_NAME,ignore.case=TRUE,
                                      invert=TRUE),]
```

```r
# Check if transactiondata has any NULL values
summary(is.na(transactiondata)) ## no columns have any na values
```

```
##     DATE          STORE_NBR       LYLTY_CARD_NBR    TXN_ID
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:246742    FALSE:246742    FALSE:246742    FALSE:246742
##   PROD_NBR       PROD_NAME        PROD_QTY       TOT_SALES
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:246742    FALSE:246742    FALSE:246742    FALSE:246742
```

```r
# print out the transaction where 200 packs of chips were bought
transactiondata[transactiondata$PROD_QTY==200,]
```

```
## # A tibble: 2 x 8
##   DATE       STORE_NBR LYLTY_CARD_NBR TXN_ID PROD_NBR PROD_NAME PROD_QTY
##   <date>         <dbl>          <dbl>  <dbl>    <dbl> <chr>        <dbl>
## 1 2018-08-19       226          226000 226201        4 Dorito C~      200
## 2 2019-05-20       226          226000 226210        4 Dorito C~      200
## # ... with 1 more variable: TOT_SALES <dbl>
```

```
# Check if this customer has had any other transactions
transactiondata[transactiondata$LYLTY_CARD_NBR==226000,] #maybe bought chips for commercial purposes
```

```
## # A tibble: 2 x 8
##   DATE       STORE_NBR LYLTY_CARD_NBR TXN_ID PROD_NBR PROD_NAME PROD_QTY
##   <date>         <dbl>          <dbl>  <dbl>    <dbl> <chr>        <dbl>
## 1 2018-08-19       226          226000 226201        4 Dorito C~      200
## 2 2019-05-20       226          226000 226210        4 Dorito C~      200
## # ... with 1 more variable: TOT_SALES <dbl>
```

```
# Remove this customer from further analysis
transactiondata<-transactiondata[transactiondata$LYLTY_CARD_NBR!=226000,]
```

```
# Summary of count by date
transaction_by_day<-transactiondata %>% group_by(DATE) %>% summarise(N=n())
```

```
# create a data frame with all the dates between 2018-07-01 and 2019-06-30
dateseq<-as.data.frame(seq(as.Date("2018-07-01"),as.Date("2019-06-30"),by="day"))
dateseq<-setNames(dateseq,"DATE")

# Find the missing date by anti_join, return all rows from dataseq where there are not matching values
anti_join(dateseq,transactiondata,by="DATE")
```

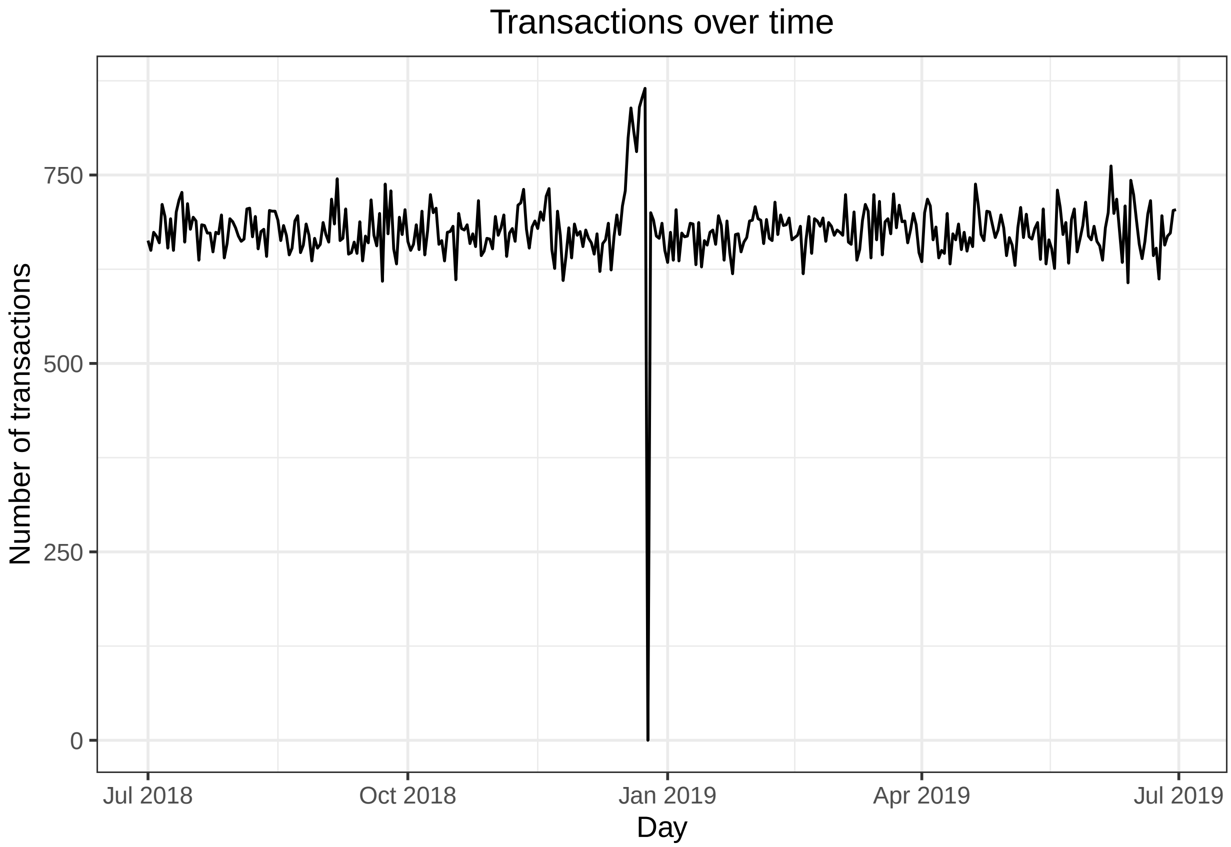
```
##         DATE
## 1 2018-12-25
```

Missing date is 2018-12-25, which is Xmas date! Store was probably closed

```
# Add this date, N=0 to transaction_by_day df
transaction_by_day<-rbind(transaction_by_day,data.frame(DATE=as.Date("2018-12-25"),N=0))
```

```
# Set theme for plots
theme_set(theme_bw())
theme_update(plot.title=element_text(hjust=0.5),plot.subtitle=element_text(hjust=0.5))
```

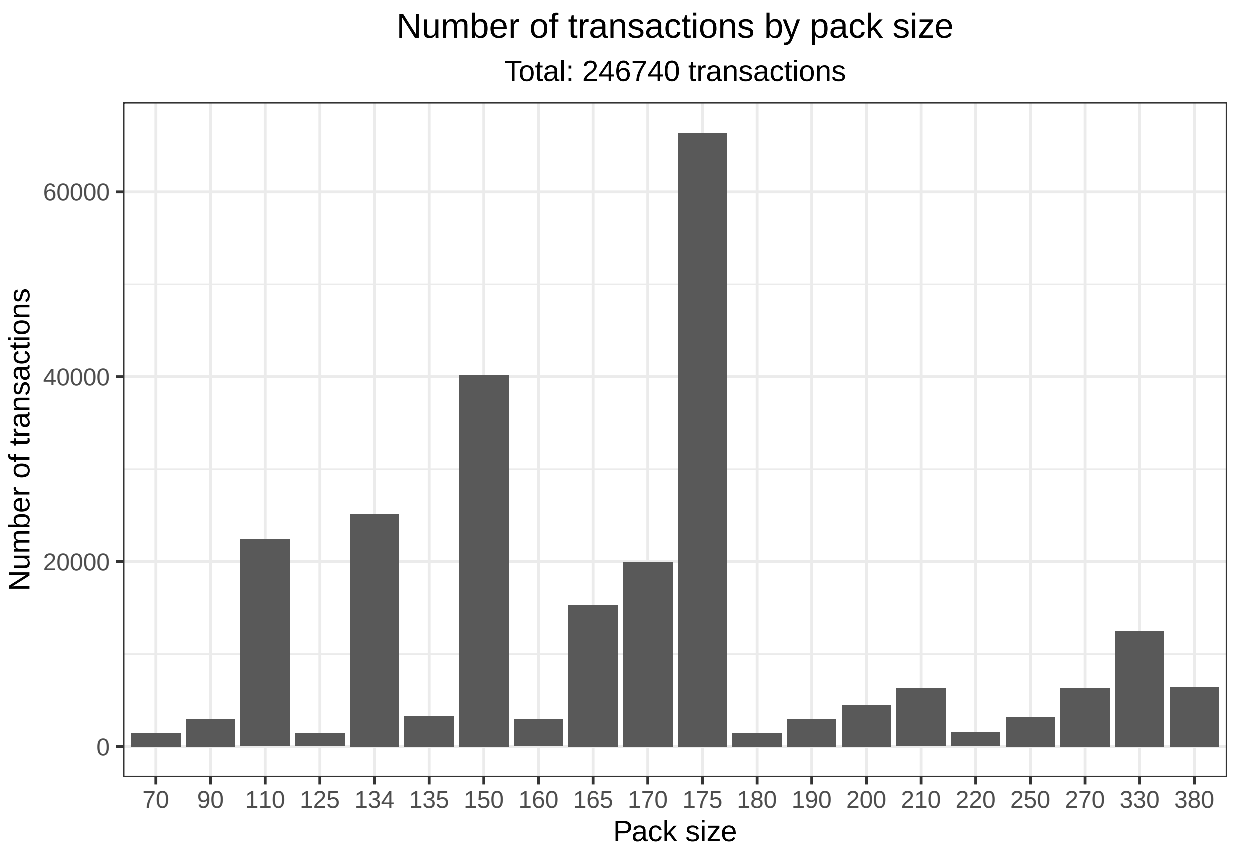
```
# line graph for transactions over time
ggplot(transaction_by_day,aes(x=DATE,y=N))+
  geom_line()+
  labs(title="Transactions over time",x="Day",y="Number of transactions")
```



Steady purchase throughout the year but higher near the end of the year

```
# add column PACK_SIZE to transactiondata
transactiondata$PACK_SIZE<- parse_number(transactiondata$PROD_NAME)
```

```
# Barplot Number of transactions ~ packsize
ggplot(transactiondata, aes(x=factor(PACK_SIZE)))+
  geom_bar()+
  labs(title="Number of transactions by pack size",x="Pack size",
       y="Number of transactions",
       subtitle=paste("Total:",nrow(transactiondata),"transactions"))
```



Pack size 175g is the most popular choice among all transactions

```
# Extract the first word starting from 1 position, to 1 position, separated by " "
transactiondata$BRAND_NAME<-stringr::word(transactiondata$PROD_NAME,1,1,sep=" ")
### Overview of all unique brand names
unique(transactiondata$BRAND_NAME)
```

```
##  [1] "Natural"    "CCs"        "Smiths"     "Kettle"     "Grain"
##  [6] "Doritos"    "Twisties"   "WW"         "Thins"      "Burger"
## [11] "NCC"        "Cheezels"   "Infzns"     "Red"        "Pringles"
## [16] "Dorito"     "Infuzions"  "Smith"      "GrnWves"    "Tyrrells"
## [21] "Cobs"       "French"     "RRD"        "Tostitos"   "Cheetos"
## [26] "Woolworths" "Snbts"      "Sunbites"
```

```
# Make some adjustments to brand names
# Find and replace Red with RRD.
transactiondata[grep("Red",transactiondata$BRAND_NAME,fixed=TRUE),"BRAND_NAME"]<-"RRD"
# Find and replace Dorito with Doritos
transactiondata[grep("Dorito",transactiondata$BRAND_NAME,fixed=TRUE),"BRAND_NAME"]<-"Doritos"
# Find and replace Infzns with Infuzions
transactiondata[grep("Infzns",transactiondata$BRAND_NAME,fixed=TRUE),"BRAND_NAME"]<-"Infuzions"
# Find and replace Snbts with Sunbites
transactiondata[grep("Snbts",transactiondata$BRAND_NAME,fixed=TRUE),"BRAND_NAME"]<-"Sunbites"
```

```
transactiondata[grep("WW",transactiondata$BRAND_NAME,fixed=TRUE),"BRAND_NAME"]<-"Woolworths"
transactiondata[grep("Grain",transactiondata$BRAND_NAME,fixed=TRUE),"BRAND_NAME"]<-"GrnWves"
transactiondata[grep("Smith",transactiondata$BRAND_NAME,fixed=TRUE),"BRAND_NAME"]<-"Smiths"
# Double check the brand names
unique(transactiondata$BRAND_NAME) #### 21 distinct brand names ####
```

```
##  [1] "Natural"    "CCs"        "Smiths"     "Kettle"     "GrnWves"
##  [6] "Doritos"    "Twisties"   "Woolworths" "Thins"      "Burger"
## [11] "NCC"        "Cheezels"   "Infuzions"  "RRD"        "Pringles"
## [16] "Tyrrells"   "Cobs"       "French"     "Tostitos"   "Cheetos"
## [21] "Sunbites"
```

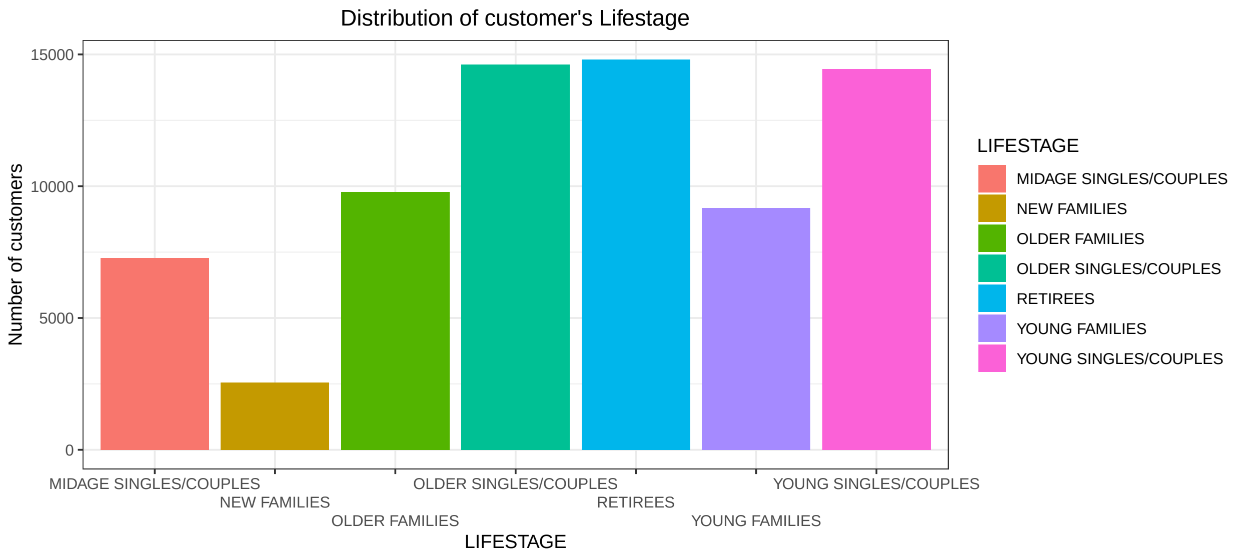
```
summary(purchasebehaviour)
```

```
##  LYLTY_CARD_NBR     LIFESTAGE          PREMIUM_CUSTOMER
##  Min.   :   1000   Length:72637       Length:72637
##  1st Qu.:  66202   Class :character   Class :character
##  Median : 134040   Mode  :character   Mode  :character
##  Mean   : 136186
##  3rd Qu.: 203375
##  Max.   :2373711
```

```
# Check if there's any NA values in any column
summary(is.na(purchasebehaviour)) # No NA
```

```
##  LYLTY_CARD_NBR  LIFESTAGE       PREMIUM_CUSTOMER
##  Mode :logical   Mode :logical   Mode :logical
##  FALSE:72637     FALSE:72637     FALSE:72637
```

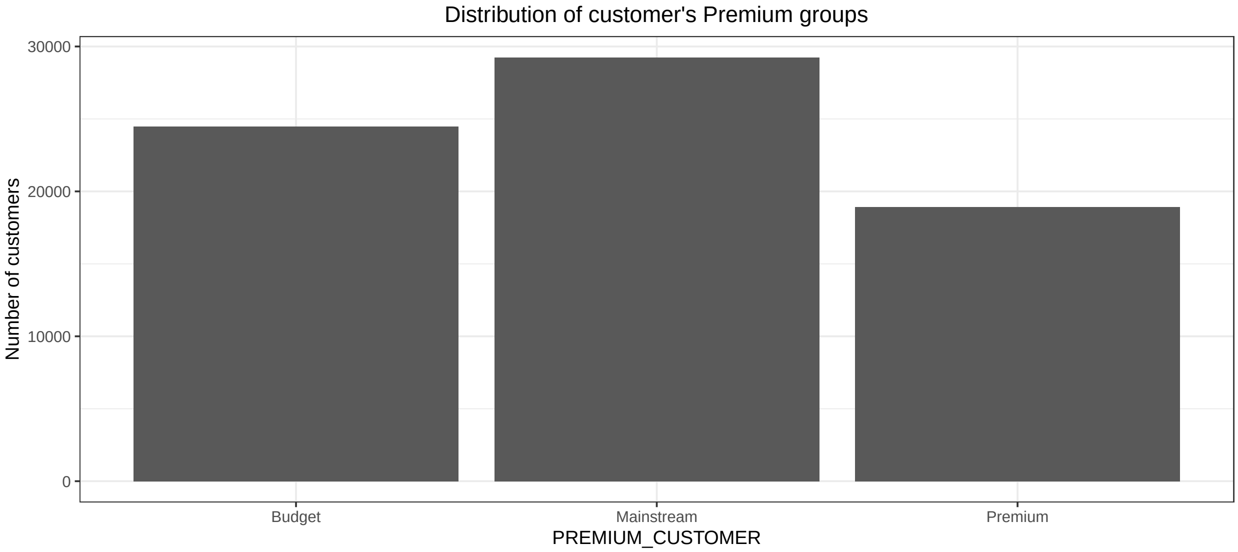
```
# Check distribution of LIFESTAGE
ggplot(purchasebehaviour,aes(x=LIFESTAGE,fill=LIFESTAGE))+
  geom_bar(show.legend=TRUE)+
  labs(title="Distribution of customer's Lifestage",y="Number of customers")+
  scale_x_discrete(guide = guide_axis(n.dodge = 3))
```



Fewer members in New Families and Midage singles/couples and Young families

Fair distribution among Retirees, Older Families and Young singles/couples

```
# Check distribution of PREMIUM_CUSTOMER
 ggplot(purchasebehaviour,aes(x=PREMIUM_CUSTOMER))+
  geom_bar()+
  labs(title="Distribution of customer's Premium groups",y="Number of customers")
```



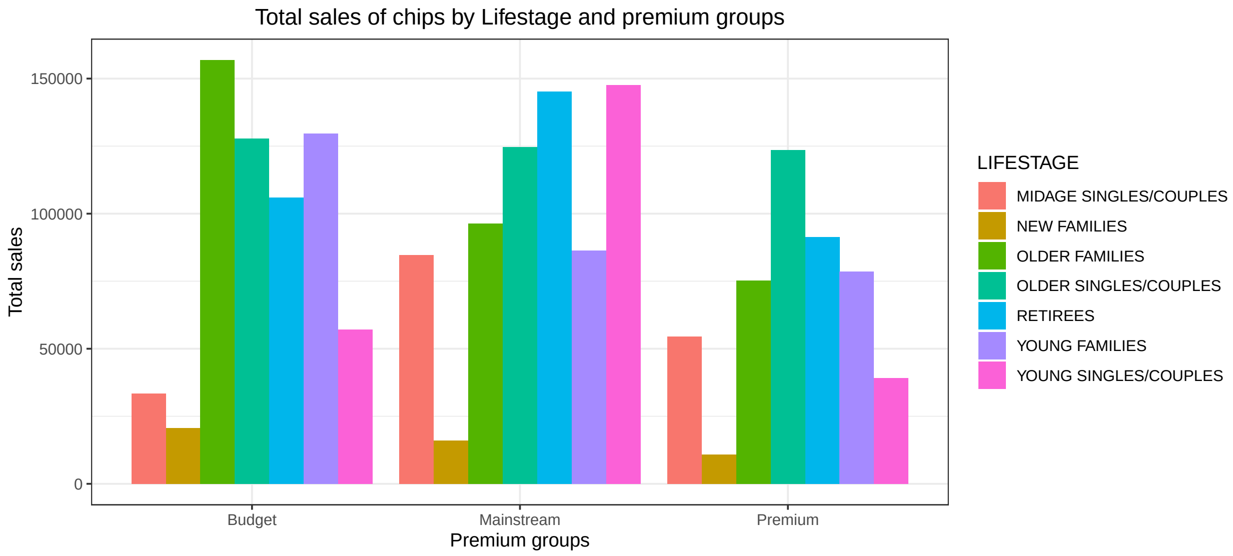
Fewer members in Premium group. Highest number of members in Mainstream.

```
# Merge transaction data to customer data
data<-left_join(transactiondata,purchasebehaviour)
```

```
summary(is.na(data$LIFESTAGE))### no. of FALSE= number of rows, so no missing customer details
```

```
##    Mode   FALSE
## logical  246740
```

```
### TOT_SALES by LIFESTAGE and PREMIUM_CUSTOMER

## use geom_bar(), with weight aes to represent the sum of sales in each group.
ggplot(data,aes(x=PREMIUM_CUSTOMER,fill=LIFESTAGE))+
  geom_bar(aes(weight=TOT_SALES),position="dodge")+
  labs(title="Total sales of chips by Lifestage and premium groups",
       y="Total sales",x="Premium groups")
```

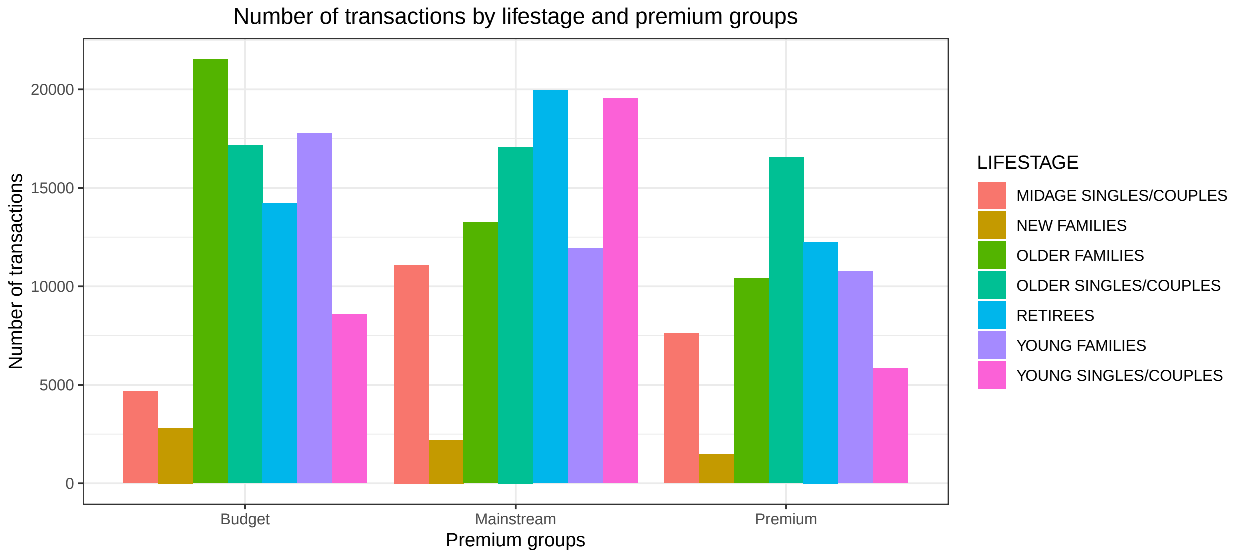


Sales are coming from Budget-Older families, Mainstream-Young singles/couples and Mainstream-retirees

Overall Premium group spend less in total

This might be subjected to the imbalance between the number of customers in each group

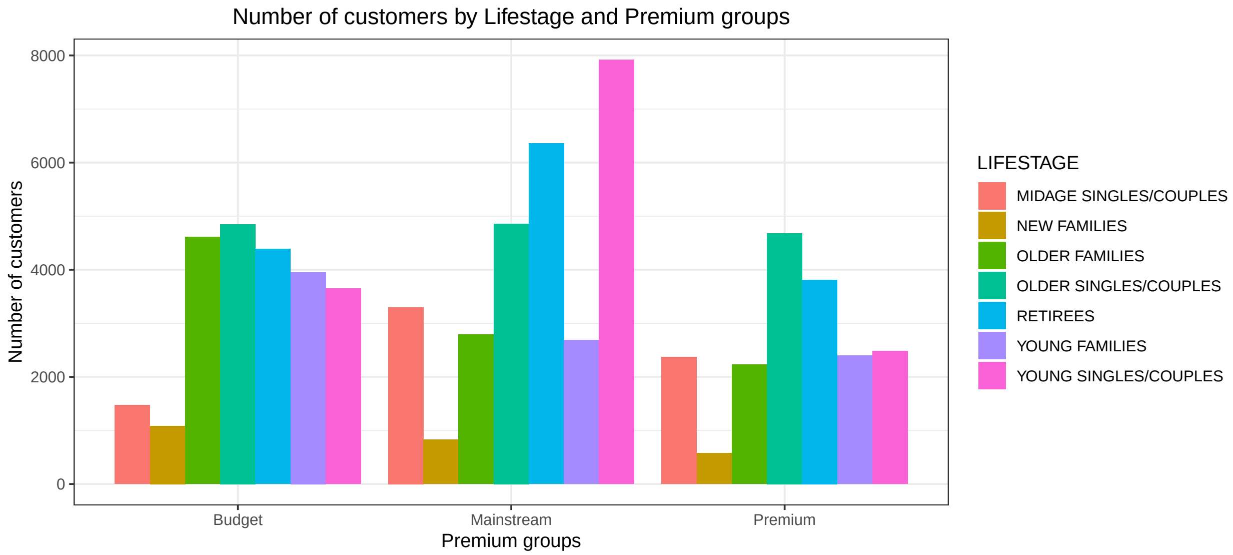
```r
# Number of transactions by Lifestage and Premium_customer
ggplot(data,aes(x=PREMIUM_CUSTOMER,fill=LIFESTAGE))+
  geom_bar(position="dodge")+
  labs(title="Number of transactions by lifestage and premium groups",
       y="Number of transactions",x="Premium groups")
```



Budget-Older families does have the highest number of transactions over the year, followed by Mainstream-Retirees and Mainstream-Young singles/couples

This might account for higher sales in these groups, but it could be because there are more members in these groups to start with, or the value of their purchase is higher than other groups

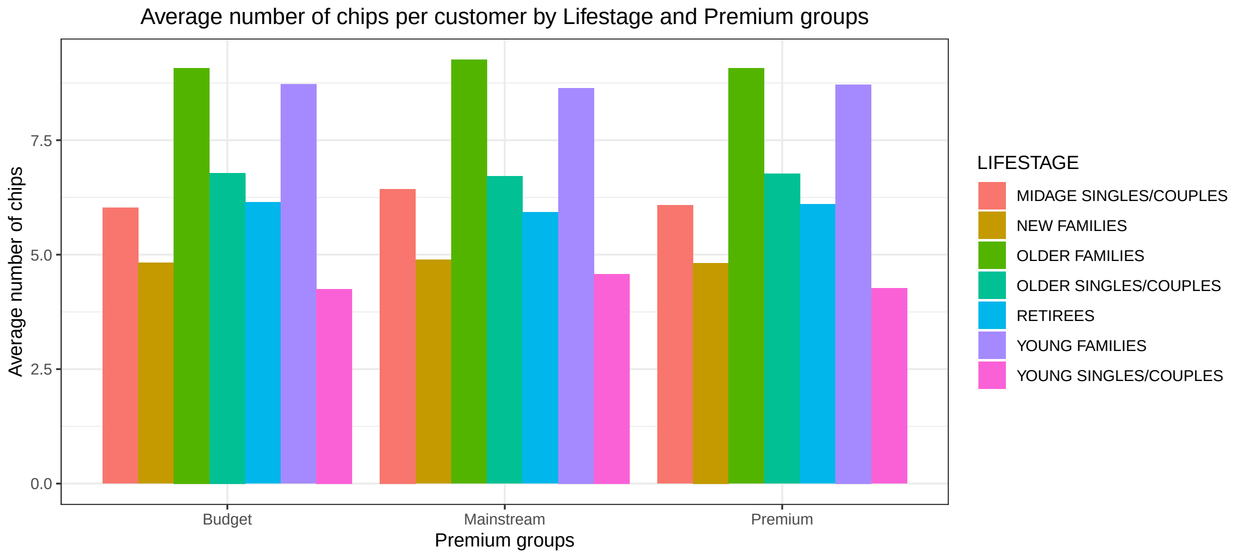
```
### Aggregate data into subsets by PREMIUM_CUSTOMER and LIFESTAGE, then apply n_distinct
#to LYLT_CARD_NBR to count distinct customers.
customer_nbr<-setNames(aggregate(data$LYLTY_CARD_NBR,
                                 by=list(data$PREMIUM_CUSTOMER,data$LIFESTAGE),
                                 FUN=n_distinct),
                       c("PREMIUM_CUSTOMER","LIFESTAGE","CUSTOMER_NUMBER"))

### Plot no. of customers ~ Lifestage and Premium_customer
ggplot(customer_nbr,aes(y=CUSTOMER_NUMBER,x=PREMIUM_CUSTOMER,fill=LIFESTAGE))+
  geom_bar(stat="identity",position="dodge")+
  labs(title="Number of customers by Lifestage and Premium groups",
       x="Premium groups",y="Number of customers")
```



Mainstream-Young singles/couples has the highest number of members (~8,000), followed by Mainstream-Retirees and Budget-Older Singles/Couples

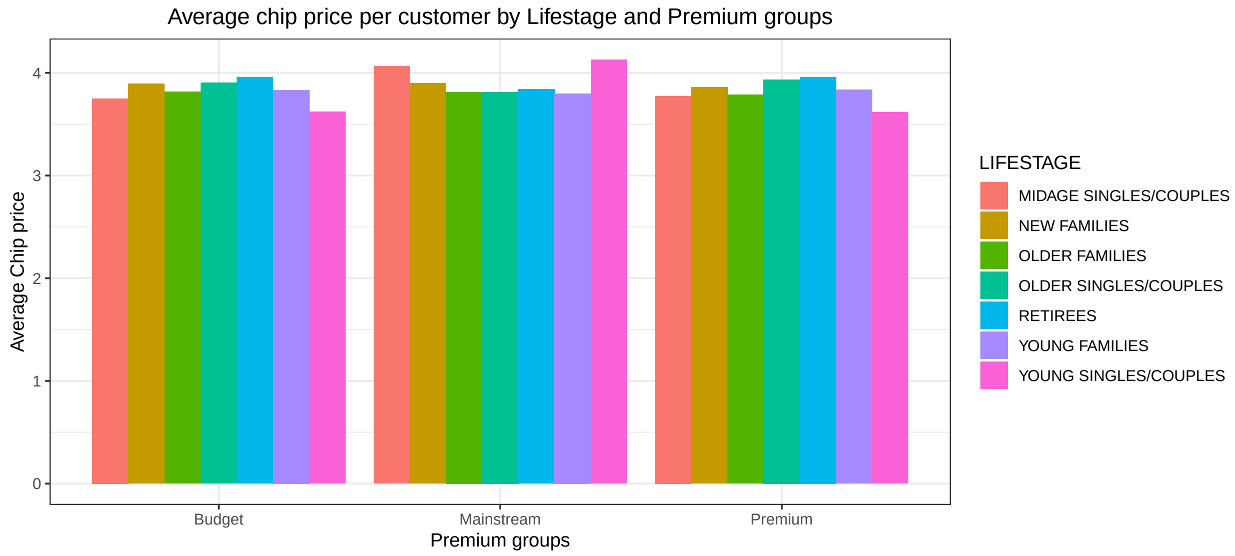
Interestingly, Budget-Older Families have the least number of members but made the highest number of transactions over the year and also contributed to the highest total sales among three segments

```
# PROD_QTY ~ PREMIUM_CUSTOMER and LIFESTAGE
# Calculate average number of PROD_QTY per customer by PREMIUM_CUSTOMER and LIFESTAGE
customer_nbr$PROD_QTY<-aggregate(data$PROD_QTY, by=list(data$PREMIUM_CUSTOMER,data$LIFESTAGE),
                                 FUN=sum)$x # Add sum of PROD_QTY by groups to customer_nbr

# Average number of chips by group
customer_nbr$AVR_PROD_QTY<-customer_nbr$PROD_QTY/customer_nbr$CUSTOMER_NUMBER

# Plot average no. of chips per customer by LIFESTAGE and PREMIUM_CUSTOMER
ggplot(customer_nbr,aes(y=AVR_PROD_QTY,x=PREMIUM_CUSTOMER,fill=LIFESTAGE))+
  geom_bar(stat="identity",position="dodge")+
  labs(title="Average number of chips per customer by Lifestage and Premium groups",
       x="Premium groups",y="Average number of chips")
```



Older families and Young families, regardless of premium groups, buy more chips per customer compared to other groups

```
# PROD_PRICE ~ PREMIUM_GROUP and LIFESTAGE
# Product_price for each unit chip per transaction
data$PROD_PRICE<-data$TOT_SALES/data$PROD_QTY

# Average price per unit chip by each customer
temp<-setNames(aggregate(data$PROD_PRICE,
                         by=list(data$PREMIUM_CUSTOMER,data$LIFESTAGE,data$LYLTY_CARD_NBR),
                         FUN=mean),
              c("PREMIUM_CUSTOMER","LIFESTAGE","LYLTY_CARD_NBR","MEAN_PRICE"))

# Calculate average price per unit chip by each customer by groups
customer_nbr$AVR_PROD_PRICE<-aggregate(temp$MEAN_PRICE,
                                       by=list(temp$PREMIUM_CUSTOMER,temp$LIFESTAGE),
                                       FUN=mean)$x

### Plot average chip price per customer by LIFESTAGE and PREMIUM_CUSTOMER
ggplot(customer_nbr,aes(y=AVR_PROD_PRICE,x=PREMIUM_CUSTOMER,fill=LIFESTAGE))+
  geom_bar(stat="identity",position="dodge")+
  labs(title="Average chip price per customer by Lifestage and Premium groups",
       x="Premium groups",y="Average Chip price")
```



Quite similar average chip price bought by each customer from different groups, but Mainstream- Young and Midage singles/couples and seem to spend more on average chip compared to the rest of the segments.

Let's do t-test to test the signicicance of PROD_PRICE purchased by these two segments in Mainstream compared to their counterparts in Budget and Premium.

```
# t-test for the difference in average chip price purchased by Young singles/couples
#between Mainstream and Budget
t.test(data=data[data$PREMIUM_CUSTOMER=="Mainstream"|data$PREMIUM_CUSTOMER=="Budget"
                 &data$LIFESTAGE=="YOUNG SINGLES/COUPLES",],
       PROD_PRICE~PREMIUM_CUSTOMER)
```

```
##
##  Welch Two Sample t-test
##
## data:  PROD_PRICE by PREMIUM_CUSTOMER
## t = -17.545, df = 10087, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2404558 -0.1921250
## sample estimates:
##     mean in group Budget mean in group Mainstream
##                 3.657366                 3.873657
```

```
#### p-value<<0.05 --> the difference is significant between Young singles/couples
#Mainstream vs Budget ####
```

```
# t-test for the difference in average chip price purchased by Young singles/couples
#between Mainstream and Premium
```

```
t.test(data=data[data$PREMIUM_CUSTOMER=="Mainstream"|data$PREMIUM_CUSTOMER=="Premium"&
                   data$LIFESTAGE=="YOUNG SINGLES/COUPLES",],
       PROD_PRICE~PREMIUM_CUSTOMER)
```

```
##
##  Welch Two Sample t-test
##
## data:  PROD_PRICE by PREMIUM_CUSTOMER
## t = 13.988, df = 6533.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1790581 0.2374279
## sample estimates:
## mean in group Mainstream    mean in group Premium
##                 3.873657                 3.665414
```

```
#### p-value<<0.05 -> the difference in chip price is significant between
#Young singles/couples Mainstream vs Premium ####
```

```
# t-test for the difference in average chip price purchased by Midage Singles/Couples
#between Mainstream and Budget
t.test(data=data[data$PREMIUM_CUSTOMER=="Mainstream"|data$PREMIUM_CUSTOMER=="Budget"&
                   data$LIFESTAGE=="MIDAGE SINGLES/COUPLES",],
       PROD_PRICE~PREMIUM_CUSTOMER)
```

```
##
##  Welch Two Sample t-test
##
## data:  PROD_PRICE by PREMIUM_CUSTOMER
## t = -8.0269, df = 5144.2, p-value = 1.229e-15
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.16215954 -0.09849827
## sample estimates:
##     mean in group Budget mean in group Mainstream
##                 3.743328                 3.873657
```

```
#### p-value<<0.05 --> the difference is significant between
#Midage singles/couples Mainstream vs Budget ####
```

```
# t-test for the difference in average chip price purchased by Midage Singles/Couples
#between Mainstream and Premium
t.test(data=data[data$PREMIUM_CUSTOMER=="Mainstream"|data$PREMIUM_CUSTOMER=="Premium"&
                   data$LIFESTAGE=="MIDAGE SINGLES/COUPLES",],
       PROD_PRICE~PREMIUM_CUSTOMER)
```

```
##
##  Welch Two Sample t-test
##
## data:  PROD_PRICE by PREMIUM_CUSTOMER
## t = 7.9204, df = 8806.2, p-value = 2.656e-15
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.07747743 0.12844051
## sample estimates:
## mean in group Mainstream    mean in group Premium
##                 3.873657                 3.770698
```

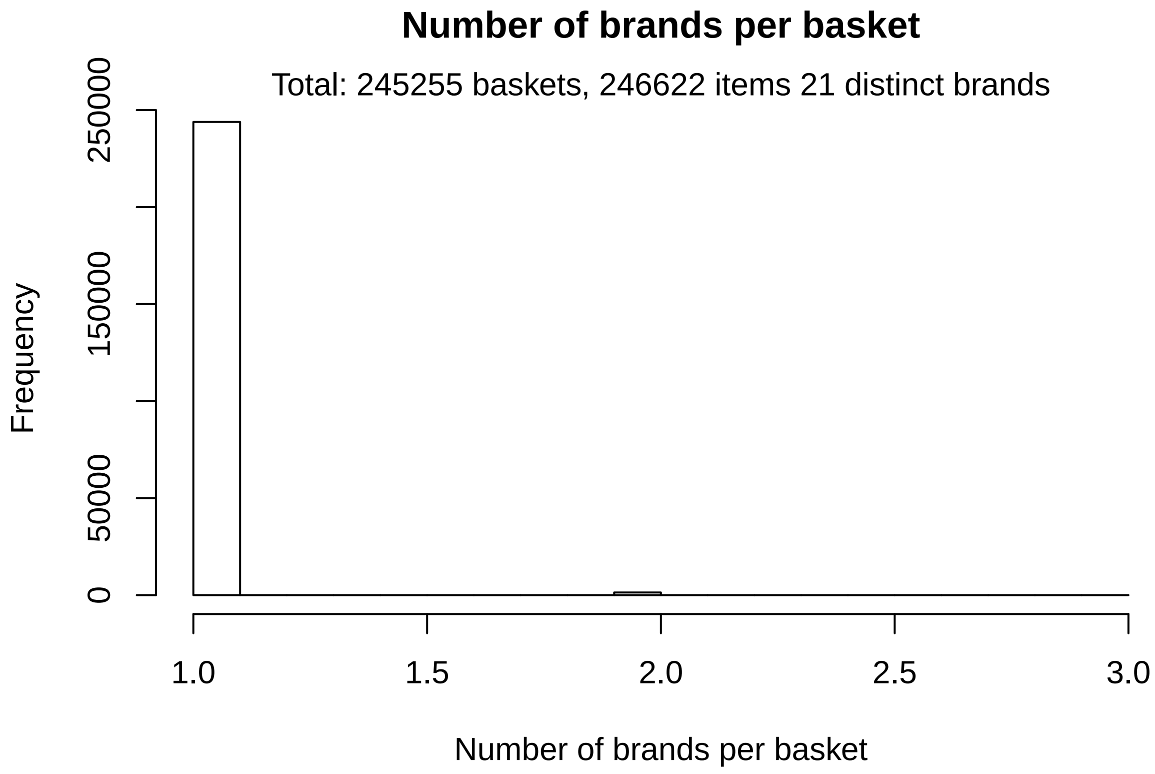
```
#### p-value<<0.05--> the difference is significant between
#Midage singles/couples Mainstream vs Premium ####
```

Overall, p-value«0.05 in four t-tests, suggesting that average chip price bought by Mainstream Young and Midage Singles/Couples is significantly higher than their counterparts in Budget and Premium group

```
## Brands preferred by each Customer Segment

# Create a copy of data to work on
data1<-data

# get the shopping baskets based on TXN_ID
Baskets<- data1 %>%
  group_by(TXN_ID) %>%
  summarise(basket=as.vector(list(BRAND_NAME)))
str(Baskets)
```

```
## tibble [245,255 x 2] (S3: tbl_df/tbl/data.frame)
##  $ TXN_ID: num [1:245255] 1 2 3 4 5 6 7 8 9 10 ...
##  $ basket:List of 245255
##   ..$ : chr "Natural"
##   ..$ : chr "RRD"
##   ..$ : chr "GrnWves"
##   ..$ : chr "Natural"
##   ..$ : chr "Woolworths"
##   ..$ : chr "Cheetos"
##   ..$ : chr "Infuzions"
##   ..$ : chr "RRD"
##   ..$ : chr "Doritos"
##   ..$ : chr "Doritos"
##   ..$ : chr "GrnWves"
##   ..$ : chr "Infuzions"
##   ..$ : chr "Smiths"
##   ..$ : chr "Doritos"
##   ..$ : chr "Kettle"
##   ..$ : chr "Doritos"
##   ..$ : chr "CCs"
##   ..$ : chr "Tostitos"
##   ..$ : chr "Kettle"
##   ..$ : chr "Kettle"
##   ..$ : chr "RRD"
##   ..$ : chr "Infuzions"
##   ..$ : chr "GrnWves"
##   ..$ : chr "Smiths"
##   ..$ : chr "Smiths"
##   ..$ : chr "GrnWves"
##   ..$ : chr "Kettle"
##   ..$ : chr "RRD"
##   ..$ : chr "Natural"
##   ..$ : chr "Smiths"
##   ..$ : chr "CCs"
##   ..$ : chr "Infuzions"
##   ..$ : chr "Smiths"
##   ..$ : chr "RRD"
##   ..$ : chr "Cobs"
##   ..$ : chr "Natural"
##   ..$ : chr "RRD"
##   ..$ : chr "Natural"
##   ..$ : chr "Burger"
##   ..$ : chr "Kettle"
##   ..$ : chr "Woolworths"
##   ..$ : chr "Smiths"
##   ..$ : chr "Thins"
##   ..$ : chr "Smiths"
##   ..$ : chr "Tyrrells"
##   ..$ : chr "Smiths"
##   ..$ : chr "Doritos"
##   ..$ : chr "Infuzions"
##   ..$ : chr "Smiths"
##   ..$ : chr "Smiths"
##   ..$ : chr "Thins"
##   ..$ : chr "Doritos"
##   ..$ : chr "Kettle"
##   ..$ : chr "Kettle"
##   ..$ : chr "Smiths"
##   ..$ : chr "Smiths"
##   ..$ : chr "Doritos"
##   ..$ : chr "Cheezels"
##   ..$ : chr "Kettle"
##   ..$ : chr "Tyrrells"
##   ..$ : chr "Twisties"
##   ..$ : chr "Doritos"
##   ..$ : chr "Thins"
##   ..$ : chr "Woolworths"
##   ..$ : chr "RRD"
##   ..$ : chr "Infuzions"
##   ..$ : chr "Smiths"
##   ..$ : chr "Infuzions"
##   ..$ : chr "GrnWves"
##   ..$ : chr "Sunbites"
##   ..$ : chr "Smiths"
##   ..$ : chr "Sunbites"
##   ..$ : chr "Kettle"
##   ..$ : chr "Smiths"
##   ..$ : chr "Sunbites"
##   ..$ : chr "Smiths"
##   ..$ : chr "Smiths"
##    ..$ : chr "Kettle"
##    ..$ : chr "Smiths"
##    ..$ : chr "Woolworths"
##    ..$ : chr "Smiths"
##    ..$ : chr "Kettle"
##    ..$ : chr "Woolworths"
##    ..$ : chr "Smiths"
##    ..$ : chr "Cobs"
##    ..$ : chr "Tostitos"
##    ..$ : chr "Natural"
##    ..$ : chr "Infuzions"
##    ..$ : chr "RRD"
##    ..$ : chr "RRD"
##    ..$ : chr "CCs"
##    ..$ : chr "Sunbites"
##    ..$ : chr "RRD"
##    ..$ : chr "Kettle"
##    ..$ : chr "Pringles"
##    ..$ : chr "Smiths"
##    ..$ : chr "Pringles"
##    ..$ : chr "French"
##    ..$ : chr "Kettle"
##    .. [list output truncated]
```

```r
# Compute transactions
transactions<-as(Baskets$basket,"transactions")
# Number of brands per basket

hist(size(transactions),main="Number of brands per basket",xlab="Number of brands per basket")
mtext(paste("Total:",length(transactions),"baskets,",sum(size(transactions)),"items",
            count(transactions@itemInfo),"distinct brands"))
```



```
## Most people only have 1 brand per transaction ##

# distribution of shoppers basket
basketSizes<-size(transactions)
summary(basketSizes)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   1.006   1.000   3.000
```

```
# quantile breakdown
quantile(basketSizes,probs=seq(0,1,0.1))
```

```
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    1    1    1    1    1    1    1    1    1    3
```

```
# get average basket amount, by TXN_ID
meanBasketAmt<-aggregate(TOT_SALES~TXN_ID,data=data1,sum)
summary(meanBasketAmt) ### 7.36 = average basket amount
```

```
##      TXN_ID           TOT_SALES
##  Min.   :      1   Min.   : 1.70
##  1st Qu.:  67558   1st Qu.: 5.80
##  Median : 135195   Median : 7.40
##  Mean   : 135136   Mean   : 7.36
##  3rd Qu.: 202678   3rd Qu.: 8.80
##  Max.   :2415841   Max.   :33.00
```

```
# get relative frequency of each brand in the transaction data
item_frequencies<-itemFrequency(transactions)

# absolute number of times a brand appear in all transactions
brandCount<-round((item_frequencies/sum(item_frequencies))*sum(basketSizes))
summary(brandCount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1418    4549    9451   11744   14198   41257
```

```
# Get top 10 brands in all transactions
orderedBrands<-sort(brandCount,decreasing=TRUE)
orderedBrands[1:10]
```

```
##     Kettle     Smiths    Doritos   Pringles        RRD  Infuzions      Thins
##      41257      30327      25204      25093      16311      14198      14072
## Woolworths       Cobs   Tostitos
##      11830       9692       9469
```

```
#how many times Kettle appears divided by total no. of transactions
orderedBrands[1]/dim(transactions)[1]
```

```
##    Kettle
## 0.1682208
```

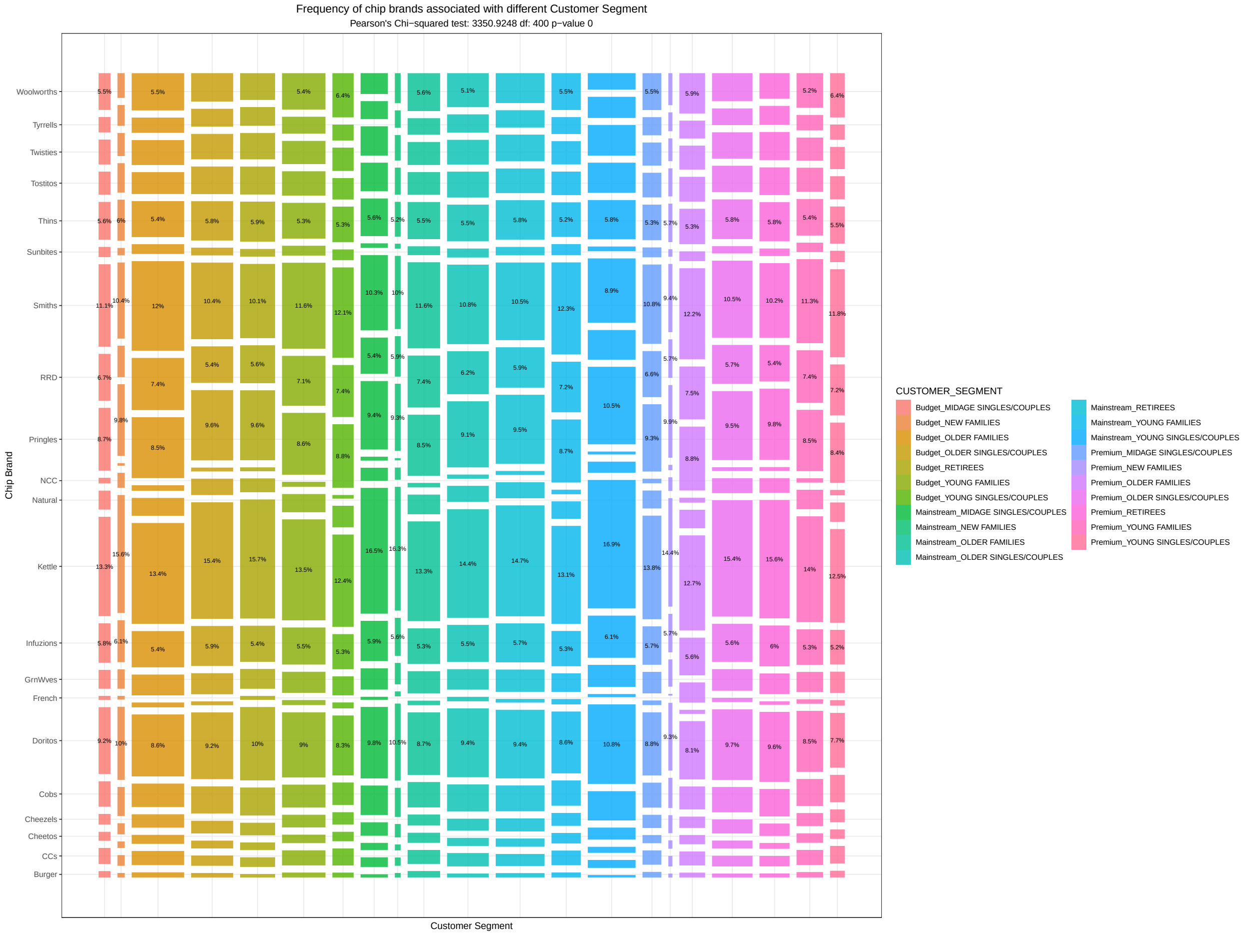
Kettle is the most popular among all customers, followed by Smiths and Doritos

The most popular brand (Kettle) appeared in their carts 16.8% of the time

```
# create customer segment column based on PREMIUM_CUSTOMER and LIFESTAGE
data1$CUSTOMER_SEGMENT<-paste(data1$PREMIUM_CUSTOMER,data1$LIFESTAGE,sep="_")

# Set CUSTOMER_SEGMENT and BRAND_NAME as categorical factors
data1$CUSTOMER_SEGMENT<-as.factor(data1$CUSTOMER_SEGMENT)
data1$BRAND_NAME<-as.factor(data1$BRAND_NAME)

# create mosaic plot
p1<-ggplot(data=data1)+
  geom_mosaic(aes(x=product(BRAND_NAME,CUSTOMER_SEGMENT),fill=CUSTOMER_SEGMENT))

# display percentage of conditional frequencies, where BRAND_NAME occurs for each CUSTOMER_SEGMENT
p1d<-ggplot_build(p1)$data %>% as.data.frame() %>% filter(.wt>0)

# function to extract percentage of conditional frequencies from mosaic plot data
compt_perc=function(x){
  d=c(x,1)-c(0,x)
  d[-length(d)]
```

```r
}

# compute conditional percentage
x=tapply(p1d$ymax,factor(p1d$fill,levels=unique(p1d$fill)),compt_perc)
x=unlist(x)
p1d$percentage=paste0(round(100*x,1),"%")

# finalize the mosaic plot
p2<-p1+
    geom_text(data=p1d,aes(x=(xmin+xmax)/2,
                                  y=(ymin+ymax)/2,
                           label=ifelse(parse_number(percentage)>5,percentage,'')),
              size=2.5) +
  scale_x_productlist(labels=NULL)+
  labs(x="Customer Segment",y="Chip Brand")

# Add Pearson Chi-square test to see the significance between chip brands and customer segment
chisq=chisq.test(xtabs(~BRAND_NAME+CUSTOMER_SEGMENT,data=data1))
subtitle=paste("Pearson's Chi-squared test:",round(chisq[[1]],4),"df:",chisq[["parameter"]][["df"]],
               "p-value",chisq[[3]])

# final graph
p2<-p2+ labs(title="Frequency of chip brands associated with different Customer Segment",
         subtitle=subtitle) +theme(axis.ticks.x = element_blank(),
                                   plot.title=element_text(hjust=0.5),
                                   plot.subtitle = element_text(hjust=0.5))
```

```r
p2
```



We can see that Chip brand is significantly associated with Customer Segment.

16.9% of Mainstream-Young singles/couples purchased Kettle, the next popular brand among this segment is Doritos and Pringles at 10.8% and 10.5% respectively. The least favorite brands are Burger, Cheetos and CCs.

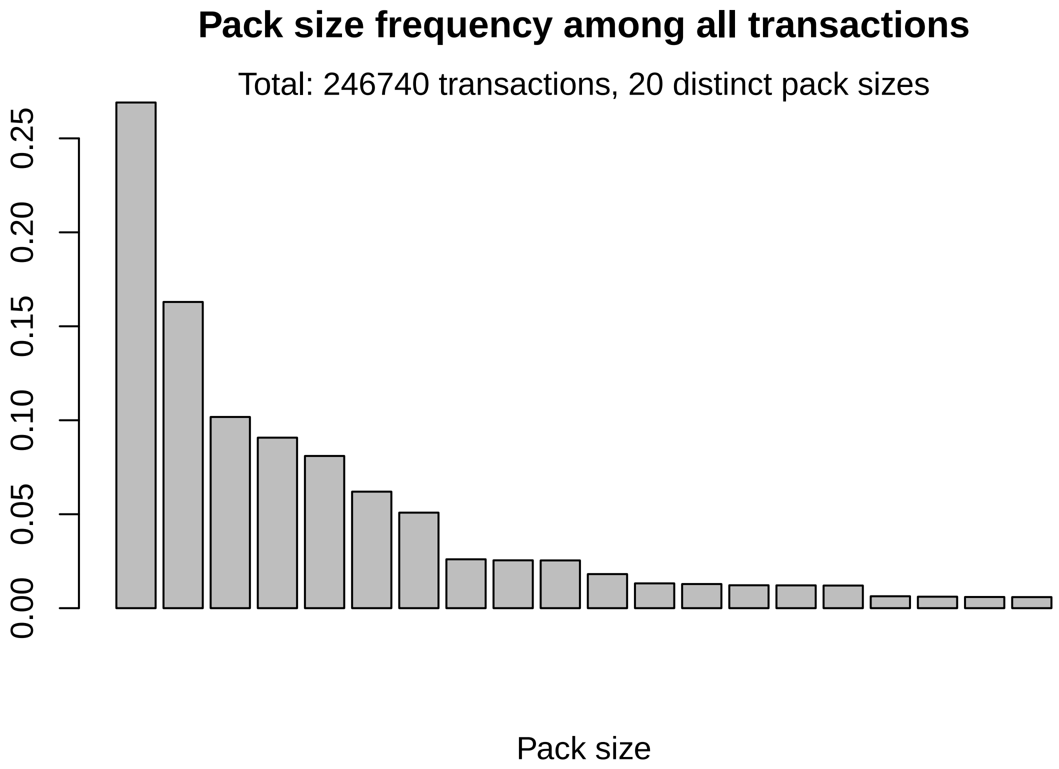
16.5% of Mainstream-Midage singles/couples purchased Kettle, the next popular brand among this segment is Smiths (10.3%), Doritos (9.8%) and Pringles (9.4%)

```r
# Set PACK_SIZE as a factor
data1$PACK_SIZE<-as.factor(data1$PACK_SIZE)
levels(data1$PACK_SIZE)
```

```
##  [1] "70"  "90"  "110" "125" "134" "135" "150" "160" "165" "170" "175" "180"
## [13] "190" "200" "210" "220" "250" "270" "330" "380"
```

```r
# Percentage of each pack size's occurrence in the whole population
packsize_freq=as.data.frame(prop.table(table(data1$PACK_SIZE)))

# Sort df according to decreasing Frequency
```

```
packsize_freq<-packsize_freq[order(-packsize_freq$Freq),]

# Top 10 pack size purchased by the whole population
barplot(packsize_freq$Freq,main = "Pack size frequency among all transactions",xlab="Pack size",
        names.arg =packsize_freq$PACK_SIZE)
mtext(paste("Total:",sum(table(data1$PACK_SIZE)),"transactions,",
            nlevels(data1$PACK_SIZE),"distinct pack sizes"))
```



Top popular pack size among all transactions are 175g, followed by 150g,134g and 110g.

```
# Create mosaic plot PACK_SIZE ~ CUSTOMER_SEGMENT
p3<-ggplot(data=data1)+
  geom_mosaic(aes(x=product(PACK_SIZE,CUSTOMER_SEGMENT),fill=CUSTOMER_SEGMENT))

# display percentage of conditional frequencies, where PACK_SIZE occurs for each CUSTOMER_SEGMENT
p3d<-ggplot_build(p3)$data %>% as.data.frame() %>% filter(.wt>0)

# compute conditional percentage
x=tapply(p3d$ymax,factor(p3d$fill,levels=unique(p3d$fill)),compt_perc)
x=unlist(x)
p3d$percentage=paste0(round(100*x,1),"%")
# Look at the distribution of percentage
```

```
summary(parse_number(p3d$percentage)) # Mean percentage is 5.0,
```

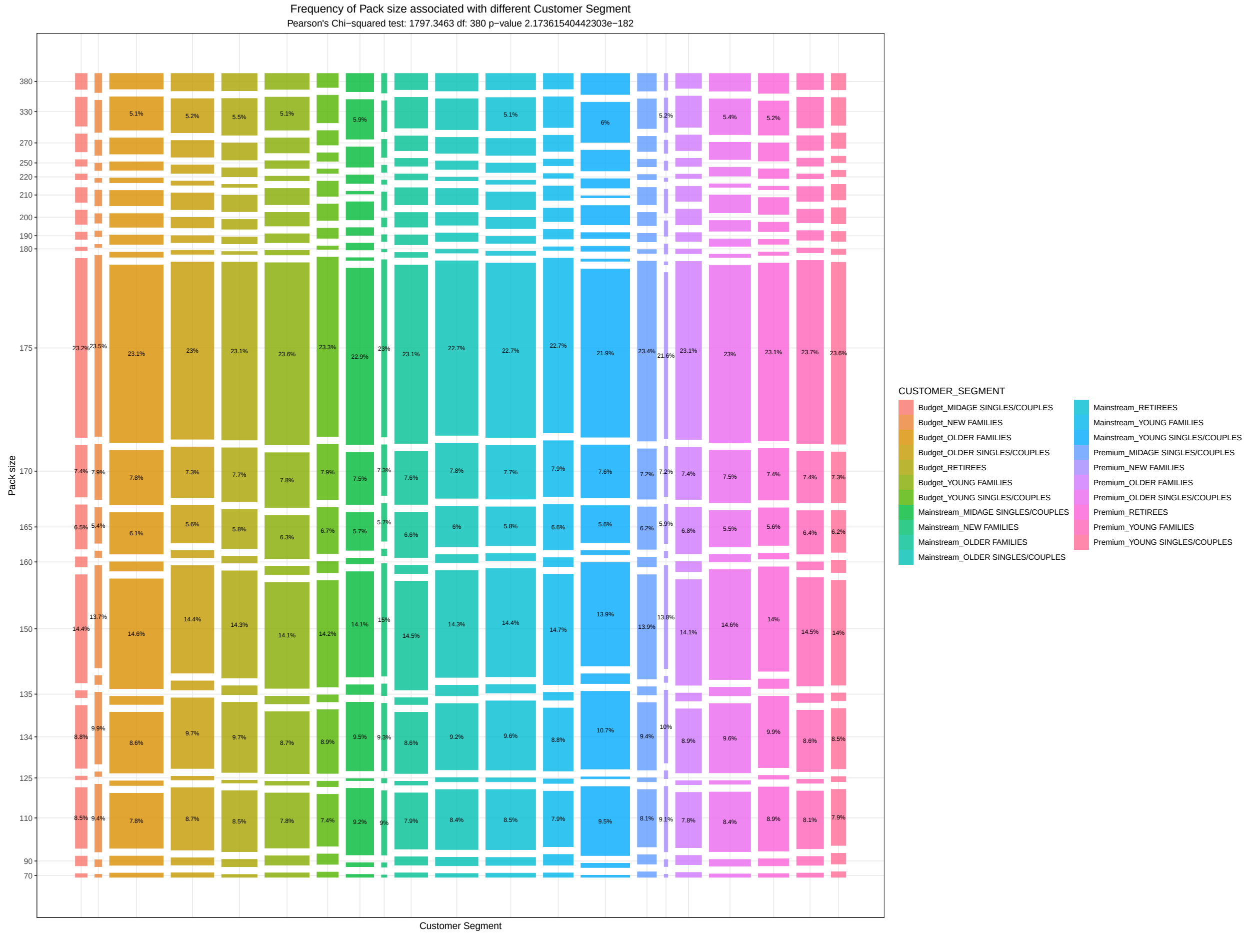
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.3     1.8     2.8     5.0     6.9    23.7
```

```
#so we’ll display only those >5 in final mosaic plot

# finalize the mosaic plot
p4<-p3+
    geom_text(data=p3d,aes(x=(xmin+xmax)/2,
                                  y=(ymin+ymax)/2,label=ifelse(parse_number(percentage)>5,percentage,’’
              size=2.5) +
  scale_x_productlist(labels=NULL)+
  labs(x="Customer Segment",y="Pack size")

# Add Pearson Chi-square test to see the significance between chip brands and customer segment
chisq_p4=chisq.test(xtabs(~PACK_SIZE+CUSTOMER_SEGMENT,data=data1))
chisq_p4 # There’s significant association between pack size and customer segments.
```

```
##
##  Pearson’s Chi-squared test
##
## data:  xtabs(~PACK_SIZE + CUSTOMER_SEGMENT, data = data1)
## X-squared = 1797.3, df = 380, p-value < 2.2e-16
```

```
subtitle_p4=paste("Pearson’s Chi-squared test:",round(chisq_p4[[1]],4),"df:",
                  chisq_p4[["parameter"]][["df"]],"p-value",chisq_p4[[3]])

# final graph
p4<-p4+ labs(title="Frequency of Pack size associated with different Customer Segment",
             subtitle=subtitle_p4) +theme(axis.ticks.x = element_blank(),
                                          plot.title=element_text(hjust=0.5),
                                          plot.subtitle = element_text(hjust=0.5))
```

```
p4
```



We can see that pack size and customer segment are significantly associated.

Mainstream Young singles/couples preferred 175g the most at 21.9%, followed by 150g at 13.9%, both of which are slightly lesser than the population average which is 26.9% (175g) and 16.3% for 150g size.

Mainstream Midage singles/couples also preferred 175g and 150g pack size.